

Preparation phase

24



User defines the following:

10

Web page content types that  
the method must recognize

N (Company News)  
C (Contact information)  
P (Product information)  
M (Management team)  
D (Company description)  
...etc...

Set of tests that provide evidence  
about the content type

15

T1 = "Number of external links on page > 5"  
T2 = "Number of internal links > 10"  
T3 = "Link text contains contact keywords  
(e.g. address, location, contact, etc)"  
T4 = "Number of people names in page > 3"  
T5 = "Page contains stock ticker symbol"  
T6 = "Page contains header starting with word  
'About...'"

...etc...

Fig. 1

Training phase

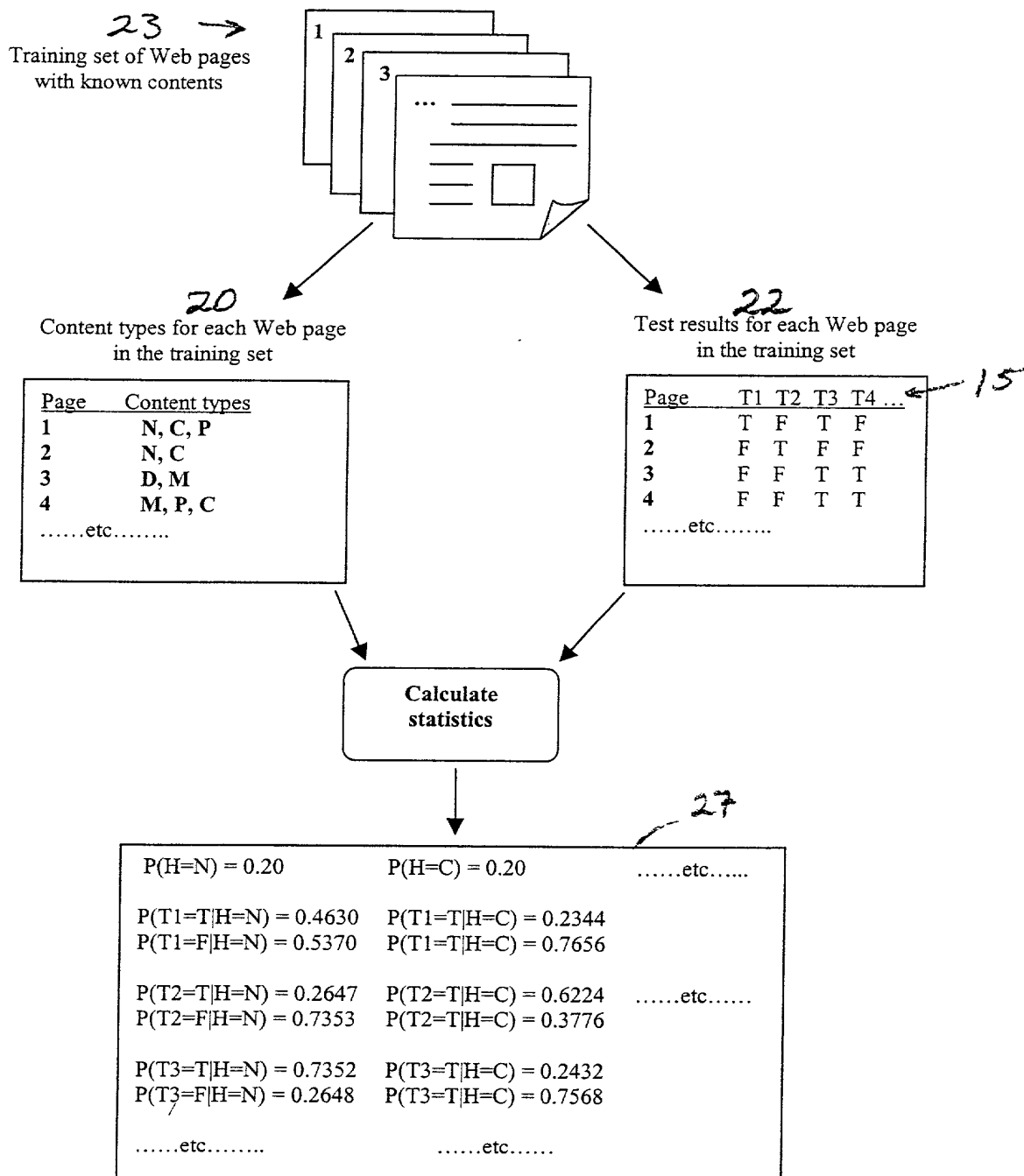


Fig. 2

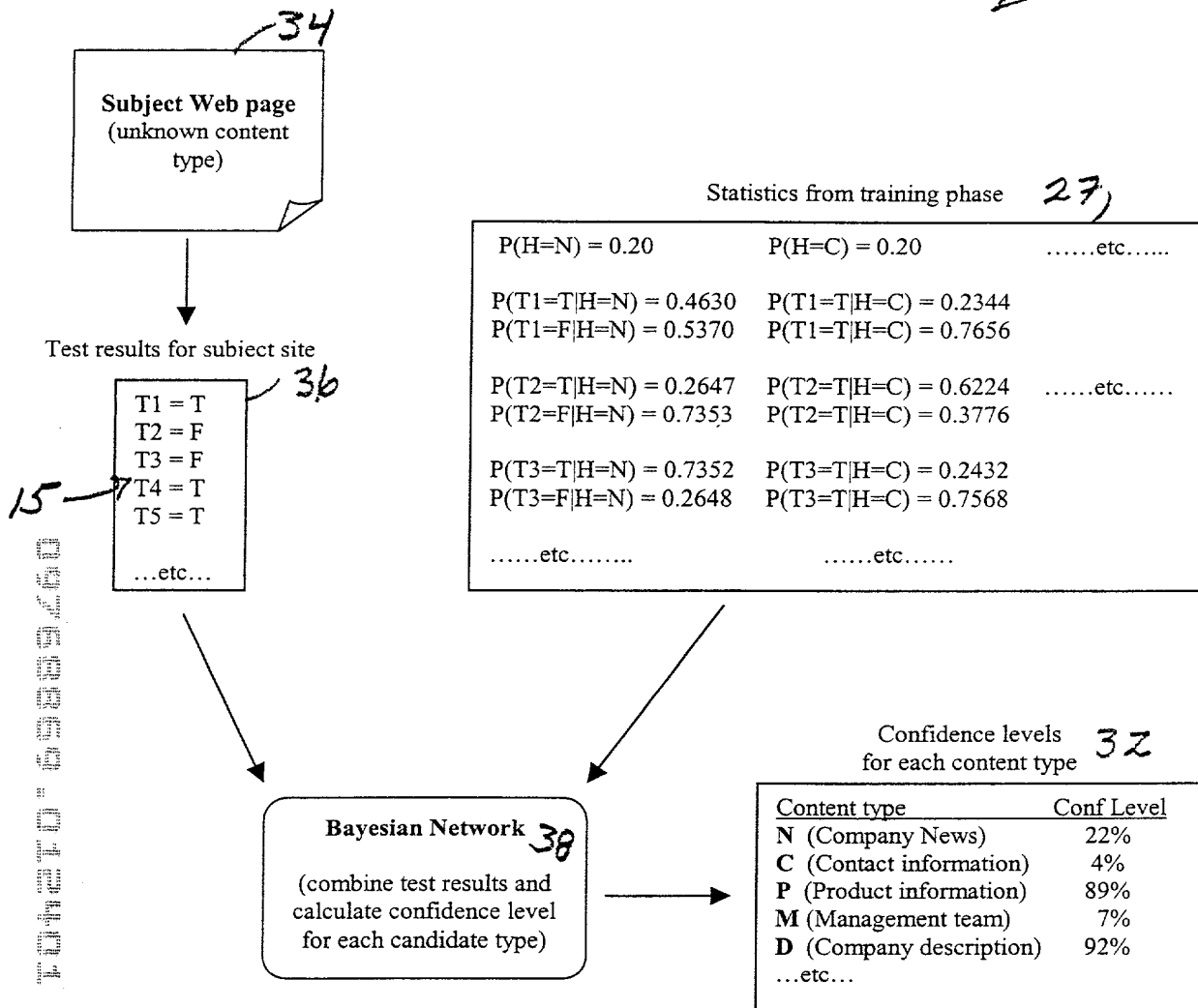
Classification phase

Fig. 3

Preferred embodiment

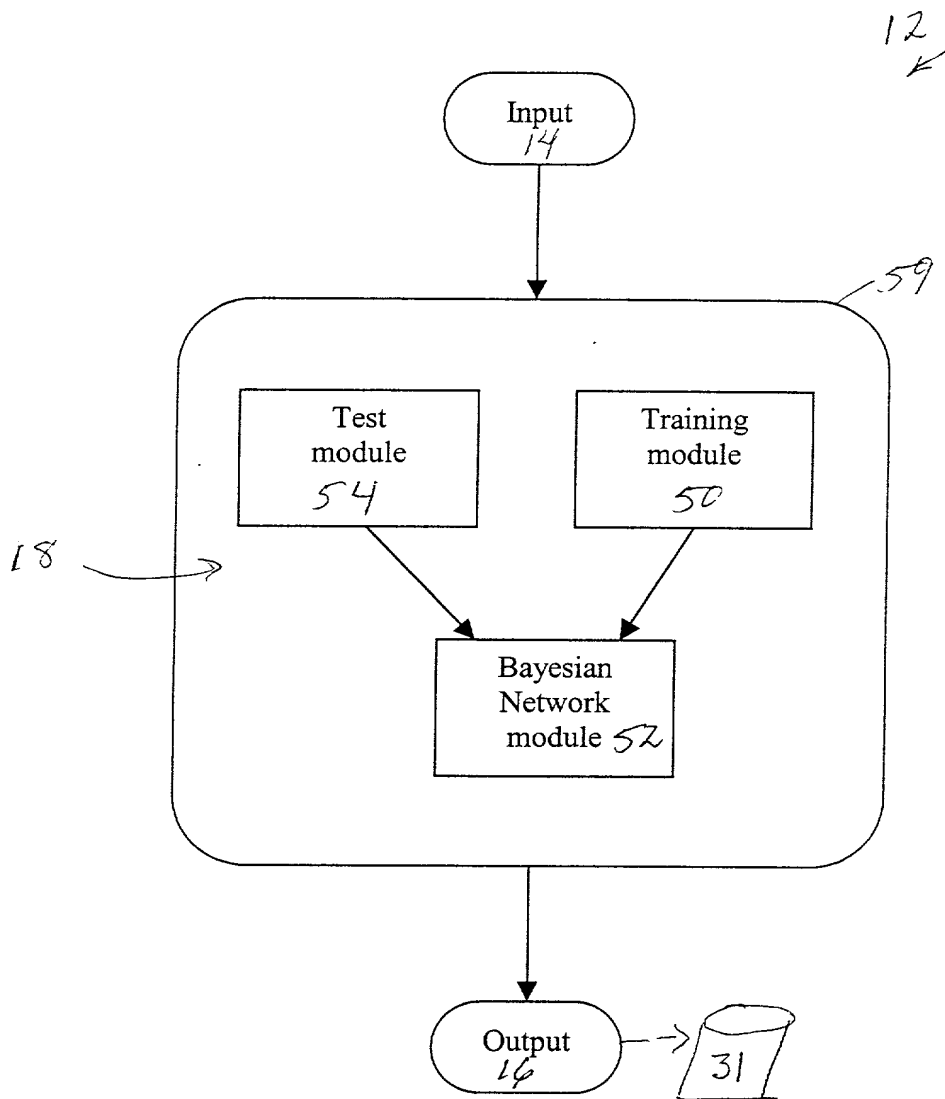


Fig. 4